IPv6 High Availability Strategies

Ivan Pepelnjak (ip@ipSpace.net)

Who is Ivan Pepelnjak (@ioshints)

- Networking engineer since 1985
- Technical director, later Chief Technology Advisor @ NIL Data Communications
- Consultant, blogger (blog.ipspace.net), book and webinar author @ ipSpace.net
- Teaching “Scalable Web Application Design” at University of Ljubljana

Focus:

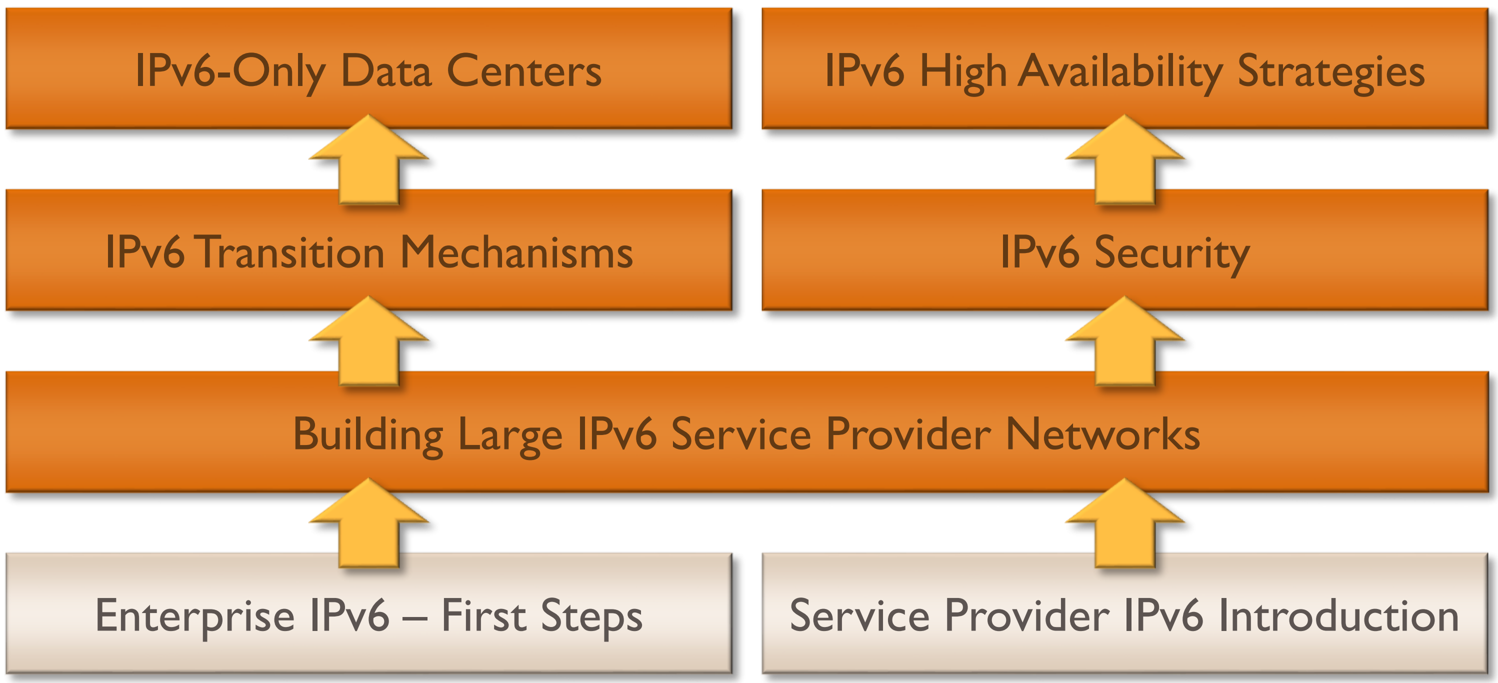
- Large-scale data centers and network virtualization
- Networking solutions for cloud computing
- Scalable application design
- Core IP routing/MPLS, IPv6, VPN

Cisco Champion

CISCO CCIE EMERITUS

vmware vEXPERT

IPv6 Webinars on ipSpace.net



Availability

- Live sessions
- Recordings of individual webinars
- Yearly subscription

Other options

- Customized webinars
- ExpertExpress
- On-site workshops

More information @ http://www.ipSpace.net/IPv6

High Availability Components

High Availability 101

A service is available = users can performs the transactions they want

Service availability includes

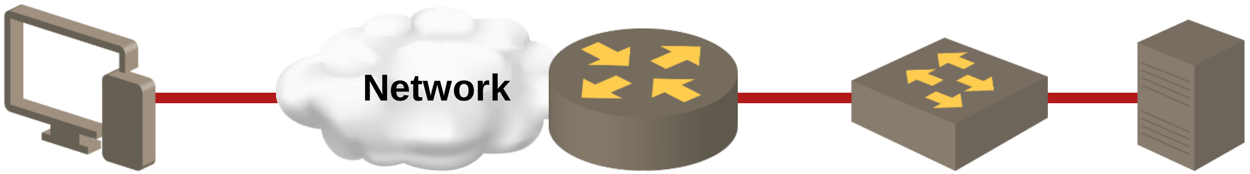
- Application availability
- Server and storage availability
- End-to-end network availability

Network availability includes

- Network services availability (DNS …)
- Network connectivity

Graceful degradation / failure resilience might be better than brute-force HA

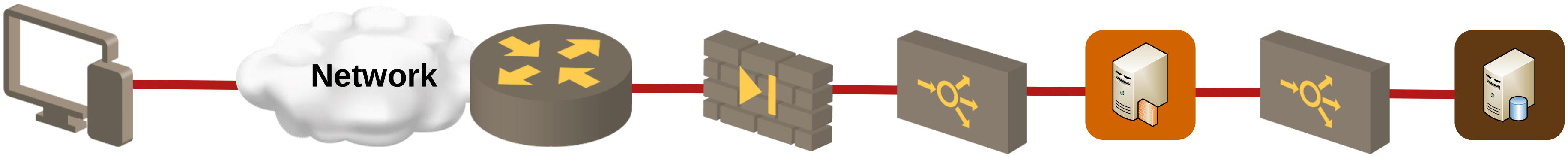
IPv6 Single-Server Applications



Network-level high availability

- Services (DNS – unchanged)
- Layer-2 (unchanged)
- First-hop router (new)
- Core network (new routing protocols, but similar)
- Multihoming (mostly unchanged, more options)

Complex IPv6 Application Stacks



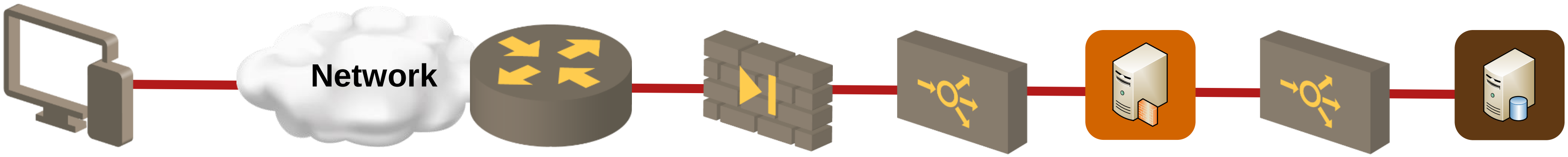
Additional application-level requirements

- Server-to-server communication
- Dependencies between application layers

Additional network-level high availability requirements

- Services: DNS, firewalls, load balancers

Beyond Networking



High availability components

- Connectivity
- Security
- Failure resilience
- Failover mechanisms
- Scale-out architectures

 © ipSpace.net 2014

Review of IPv6 First-Hop Mechanisms

Review: Configuring Host IPv6 Parameters

Minimum set of parameters:

- Host IPv6 address
- Routing information (minimum: first-hop router’s IPv6 address)
- DNS server IPv6 address (could use IPv4 DNS server in dual-stack environments)

Configuration mechanisms:

- Static configuration (servers, routers)
- Stateless Autoconfiguration (SLAAC) using Router Advertisements
- DHCPv6-based configuration

© ipSpace.net 2014

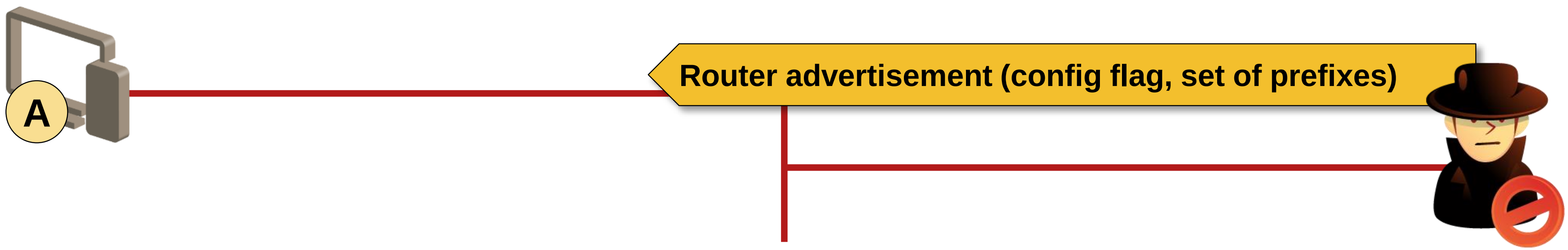
Review: Dynamic Host Configuration Options

Parameter	ICMPv6 (ND/RA)	DHCPv6
Host IPv6 address	Yes (SLAAC)	Yes
First hop router’s IPv6 address	Yes (RA)	No
DNS server’s IPv6 address	Yes (RFC 6106)	Yes

- RFC 6106 is not widely supported yet
- In most cases you need both RA and DHCPv6
- SLAAC with dynamic DNS registration is preferred to DHCPv6-based address allocation on client segments

 © ipSpace.net 2014

Why Is This Relevant?



An intruder might start sending IPv6 RA messages

- IPv6 is enabled by default on most operating systems
- Servers will auto-configure themselves
- Intruder can advertise itself as IPv6 default router and IPv6 DNS
- IPv6 DNS might take precedence over IPv4 DNS
- IPv6 transport **will** take precedence over IPv4 transport
- With proper RA messages (prefixes without on-net flag) all traffic goes through the intruder's node

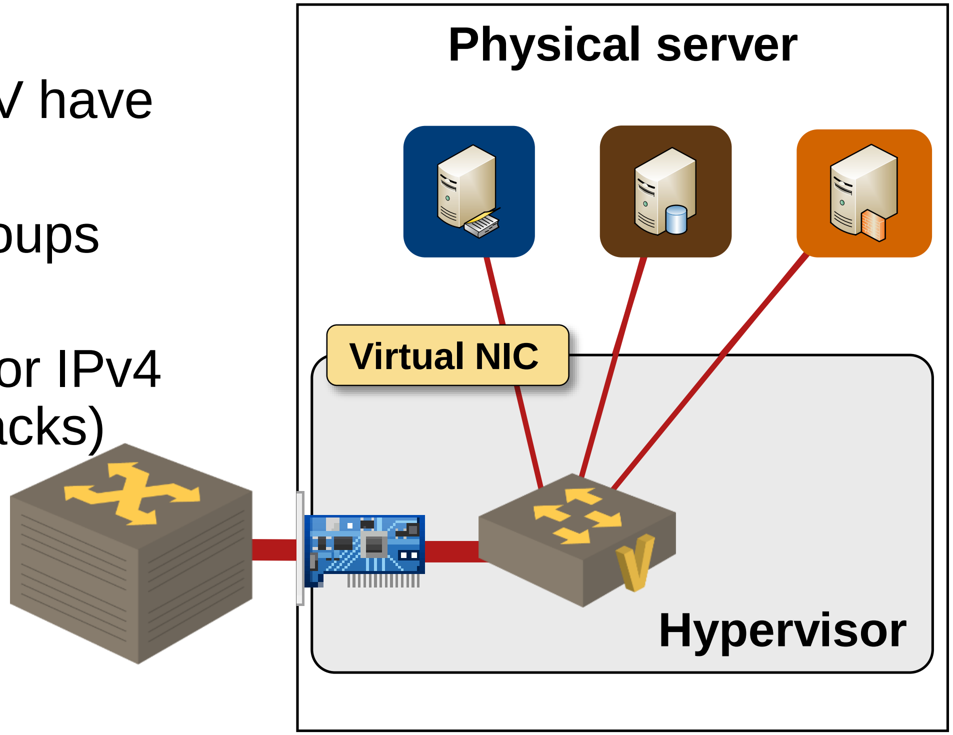
First-hop IPv6 security mechanisms are a MUST

The Virtual Fiasco

- First-hop security MUST be implemented on the first layer-2 switch
- In virtual environments the first switch is the virtual switch
- Virtual switch MUST implement IPv6 first-hop security features: RA guard, DHCPv6 guard, Source/Destination guard, Binding Integrity guard

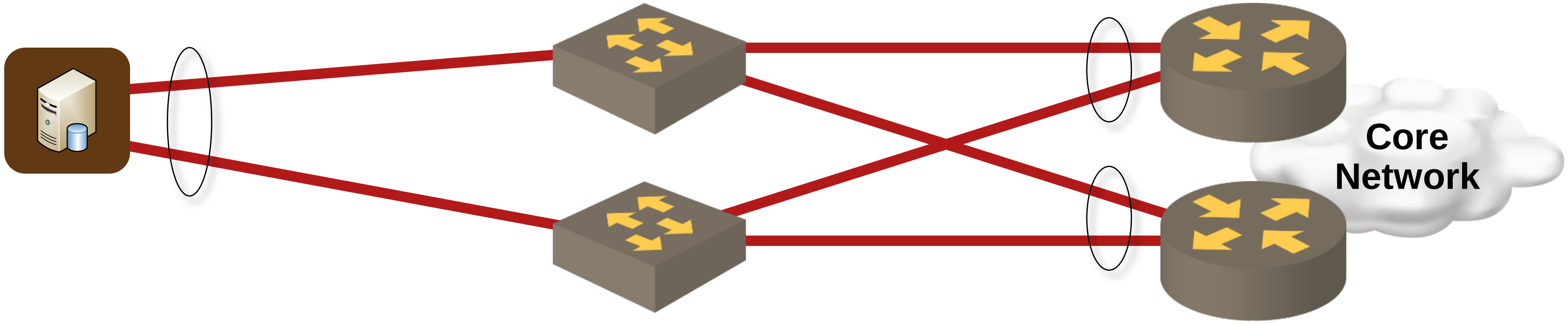
State-of-the-art:

- vSphere 5.5, vCNS 5.5 and Nexus 1000V have no IPv6 security features
- OpenStack Havana has IPv6 security groups (and little else)
- Hyper-V implements layer-3 forwarding for IPv4 and IPv6 (and thus blocks most IPv6 attacks)
- Amazon VPC does not support IPv6 (but does not propagate it either)



IPv6 First-Hop High Availability

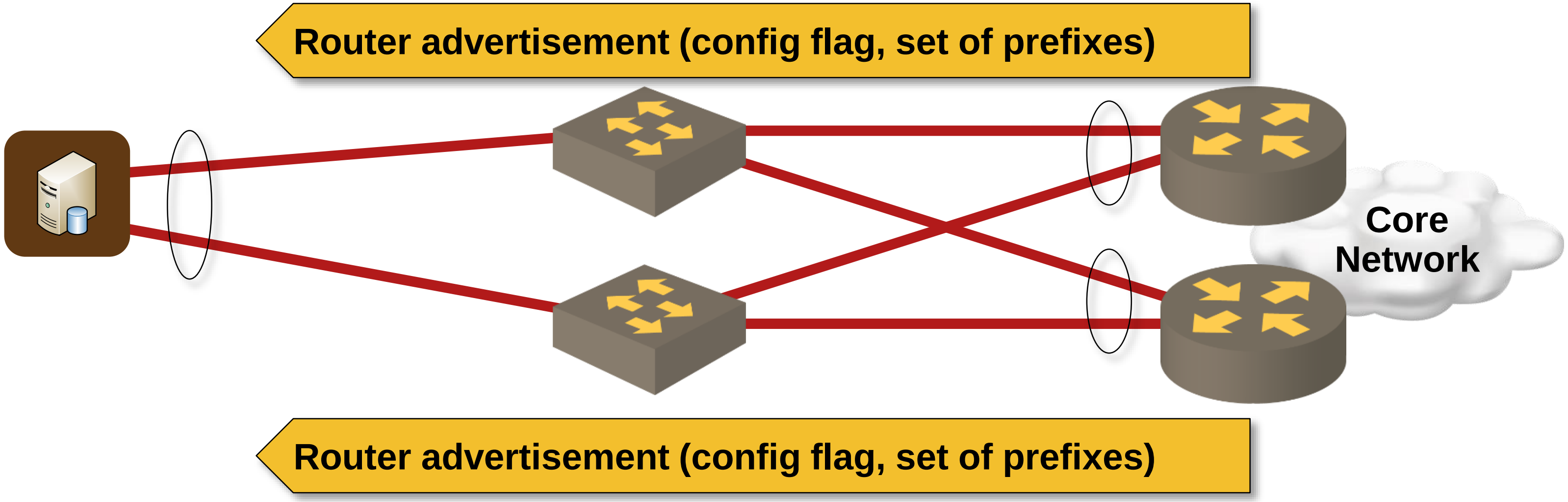
Typical High-Availability Setup



IPv6-specific modifications:

- No changes on servers (all NIC teaming modes work as expected)
- No changes on L2 switches (might need MLD snooping)
- First-hop L3 switches must be configured for high-availability environment

Router Advertisements in Dual-Router Environment



All routers advertise their presence with RA messages

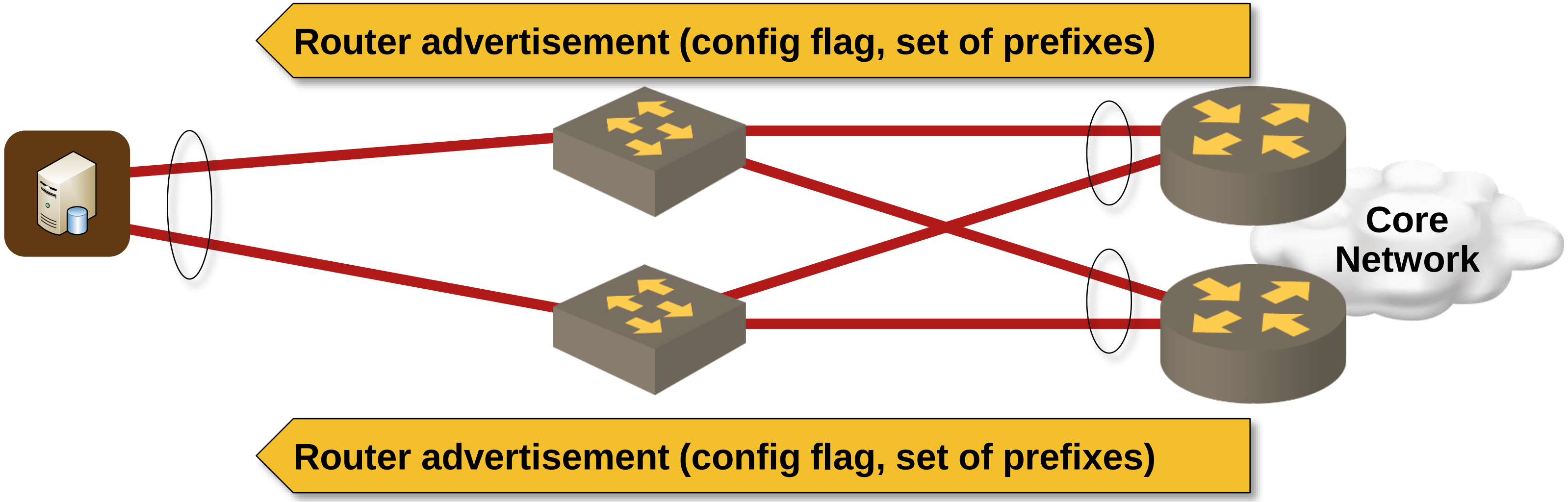
- Router's LLA and physical MAC address

Host behavior varies between operating systems (and OS versions)

- Use the first RA received as long as it's valid
- Load-balance between all first-hop routers
- Use the last RA received (flip-flopping between routers)

 © ipSpace.net 2014

Are Router Advertisements Good Enough?



RA timers can be adjusted on most routers and L3 switches

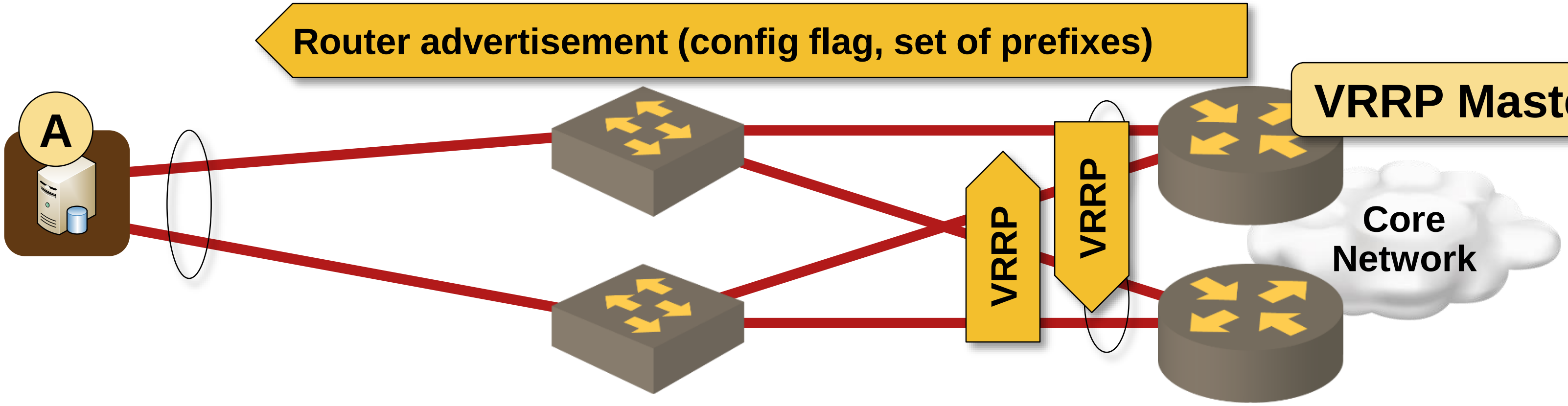
- Minimum RA interval = 30 msec (Cisco IOS)
- Minimum RA lifetime = 1 sec
- Hosts will stop using a failed router after RA expiration

RA-based failover

- Uses CPU cycles on every attached host
- Might be good enough in some environments

© ipSpace.net 2014

VRRP v3 = FHRP for IPv6

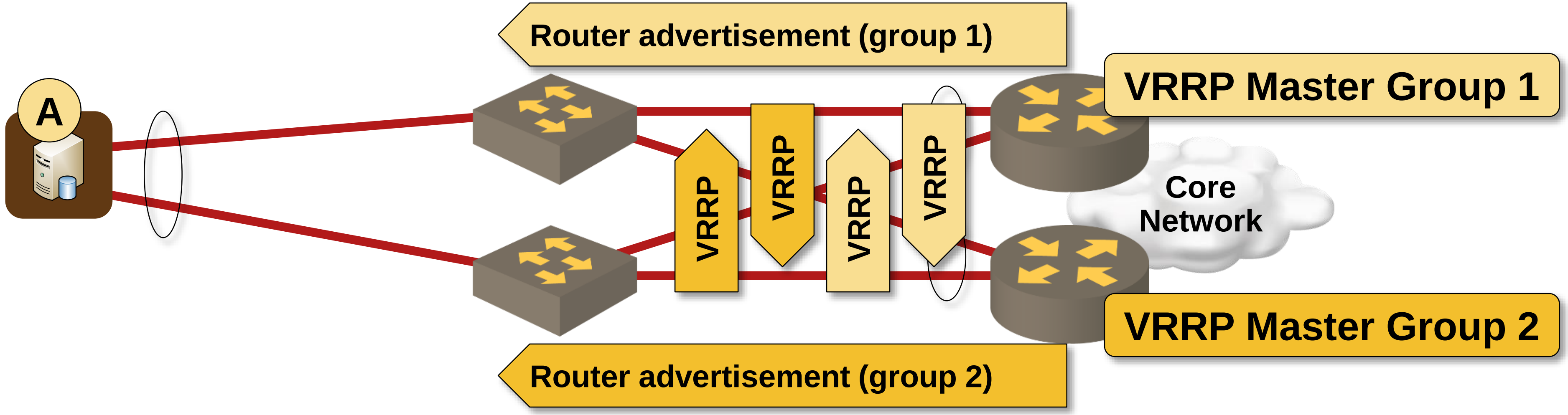


- VRRP configured on server-facing subnets
- Routers elect VRRP master
- VRRP master sends RA messages with VRRP IPv6 and VRRP MAC address
- VRRP backup router takes over VRRP MAC address after VRRP primary router failure

Sub-second convergence is possible (based on VRRP implementation)

 © ipSpace.net 2014

Load Balancing with VRRP v3



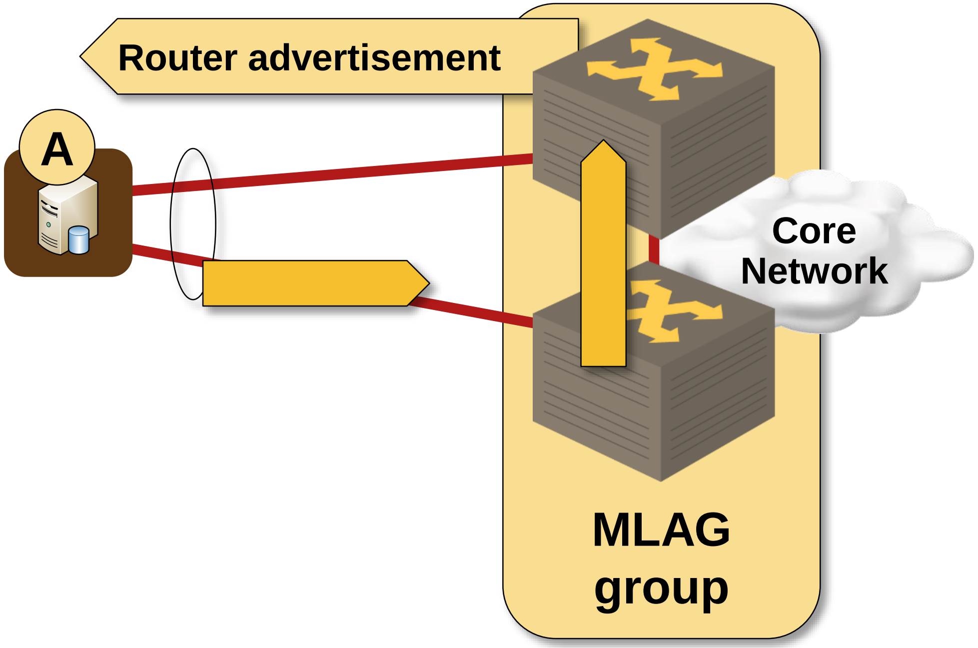
- Multiple VRRP groups configured on the same interface
- Multiple VRRP masters (one per group)
- Each VRRP master sends RA messages with its group’s IPv6 and virtual MAC address
- Hosts **might** load-balance across multiple VRRP routers

Might require static server configuration (no first-hop router in DHCPv6)

 © ipSpace.net 2014

First-Hop Redundancy on Layer-3 Switches

- Each L3 switch advertises its own physical MAC address
- Packet forwarding may become suboptimal
- Loop prevention logic might prevent proper packet forwarding



Correct design:

- Use VRRP v3 (or HSRP for IPv6)
- Both switches forward traffic sent to virtual MAC address

 © ipSpace.net 2014

Service Endpoint High Availability

IPv6 Solutions Almost Identical to IPv4 Solutions

Local high availability

- Clusters with shared IP address
- Load balancers

Redundant Internet connectivity

- BGP multihoming
- NAT/NPT with multiple uplinks (clients only)
- Mobile IP (clients only – better integrated in IPv6)
- LISP (new)

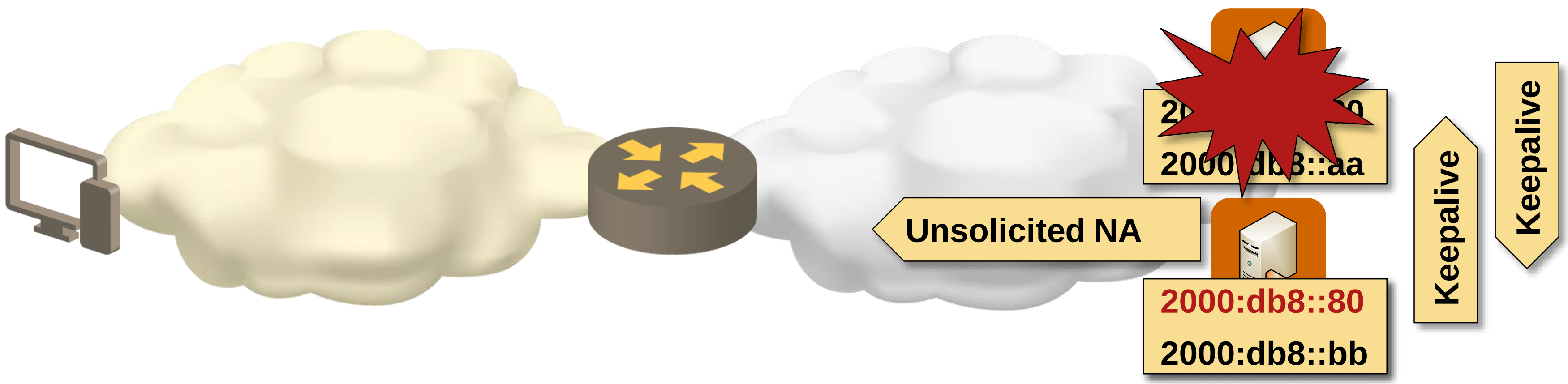
Global high-availability

- DNS-based solutions (including geolocation)
- Anycast

 © ipSpace.net 2014

Local Endpoint HA Solutions

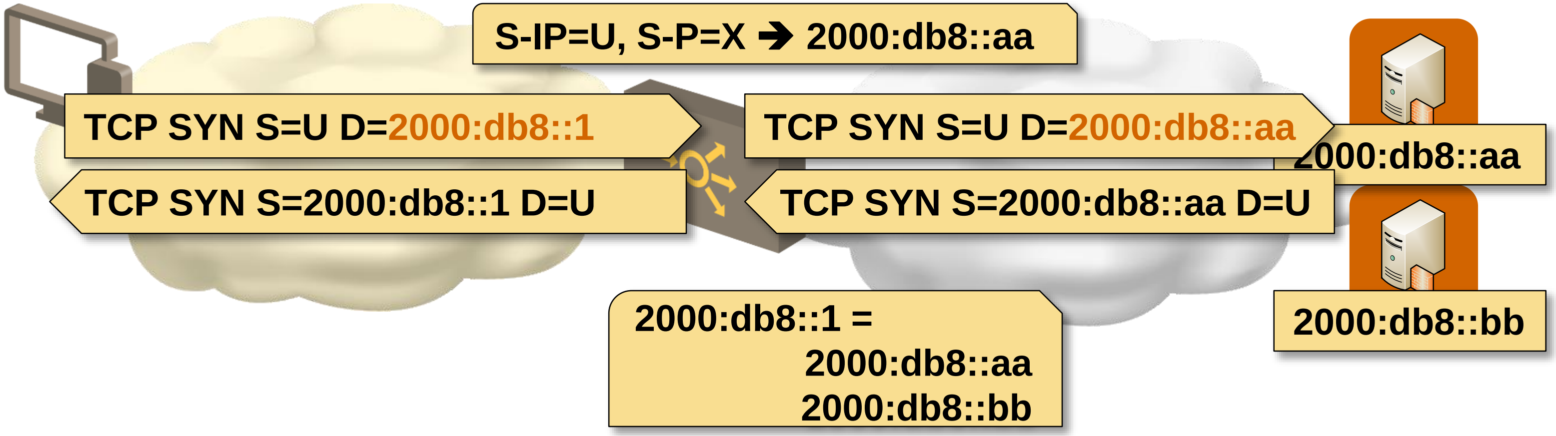
IPv6 Server Clusters



- Almost identical to IPv4 solution
- Each cluster node has a “regular” IPv6 address
- Primary node (per service) owns service IPv6 address
- Node availability checked with a keepalive protocol between cluster members
- Backup node takes over services and IPv6 addresses of a failed primary node
- Backup node sends unsolicited neighbor advertisement (equivalent to gratuitous ARP) to purge ND caches in all adjacent nodes

© ipSpace.net 2014

Load Balancers



SLB66 is almost identical to SLB44

- Load balancer in the forwarding path (destination NAT)
- SNAT for out-of-path load balancer (source + destination NAT)
- Direct server return (shared destination address, no NAT)

SLB is needed due to TCP and Socket API limitations

© ipSpace.net 2014

End-to-End High Availability

If Only TCP Stack Had Session Layer

Socket API

```
memset(&hints, 0, sizeof(hints));
hints.ai_family = PF_UNSPEC;
hints.ai_socktype = SOCK_STREAM;
error = getaddrinfo("example.com", "http", &hints, &res0);
if (error) { errx(1, "%s", gai_strerror(error)); }

s = -1;
for (res = res0; res; res = res->ai_next) {
        s = socket(res->ai_family, res->ai_socktype, res->ai_protocol);
        if (s < 0) { cause = "socket"; continue; }

        if (connect(s, res->ai_addr, res->ai_addrlen) < 0) {
                cause = "connect";
                close(s);
                s = -1;
                continue;
        }

        break;  /* okay we got one */
}
if (s < 0) { err(1, "%s", cause); }
```


© ipSpace.net 2014

Dual Stack Brokenness

	Firefox	Firefox fast-fail	Chrome	Opera	Safari	Explorer
MAC OS X 10.7.2 8.0.1	8.0.1	16.9.912.41 b	11.52	5.1.1	-	
	75s	**0ms**	**300ms**	**75s**	**270ms**	-
Windows 7	8.0.1	8.0.1	15.0.874.121 m	11.52	5.1.1	9.0.8112.16421
	21s	**0ms**	**300ms**	**21s**	**21s**	21s
Windows XP	8.0.1	8.0.1	15.0.874.121 m	11.52	5.1.1	9.0.8112.16421
	21s	**0ms**	**300ms**	**21s**	**21s**	21s
Linux 2.6.40.3-0.tc15	8.0.1	8.0.1	16.9.912.41 b	11.60 b	-	
	96s	**0ms**	**300ms**	**189s**		
iOS 5.0.1	-	-	-	-	?	-
					720ms	

Source: http://www.potaroo.net/ispcol/2011-12/esotropia.html

© ipSpace.net 2014

Dual Stack Brokenness

Traditional approach: prefer IPv6 over IPv4

- Fails miserably (after TCP timeout) in broken IPv6 environments
- No fast fallback to IPv4
- Coded in most well-written applications

Happy Eyeballs approach

- IPv4 and IPv6 sessions established (almost) in parallel
- Inherently non-deterministic
- Tests session establishment, not data flow
- PMTUD brokenness is not detected

Network services considerations

- IPv4 and IPv6 services and filters are usually configured separately

Avoid complex dual-stack environments

© ipSpace.net 2014

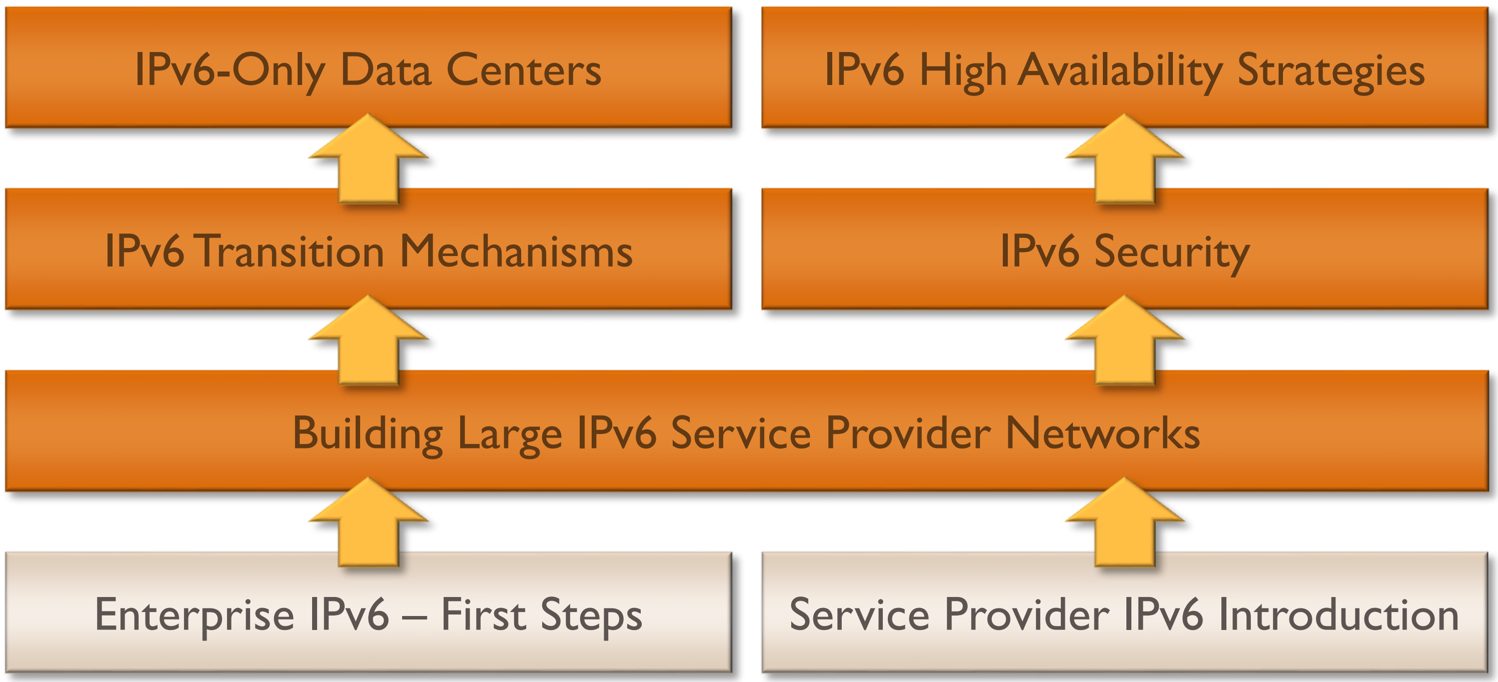
Conclusions

Conclusions

- Minor differences between IPv4 and IPv6 HA solutions
- Fundamental problems are unsolved
- Dual-stack environments with happy eyeballs are inherently non-deterministic
- Avoid the complexity of dual-stack environments whenever possible ➔ consider IPv6-only data center

© ipSpace.net 2014

IPv6 Webinars on ipSpace.net



Availability

- Live sessions
- Recordings of individual webinars
- Yearly subscription

Other options

- Customized webinars
- ExpertExpress
- On-site workshops

More information @ http://www.ipSpace.net/IPv6

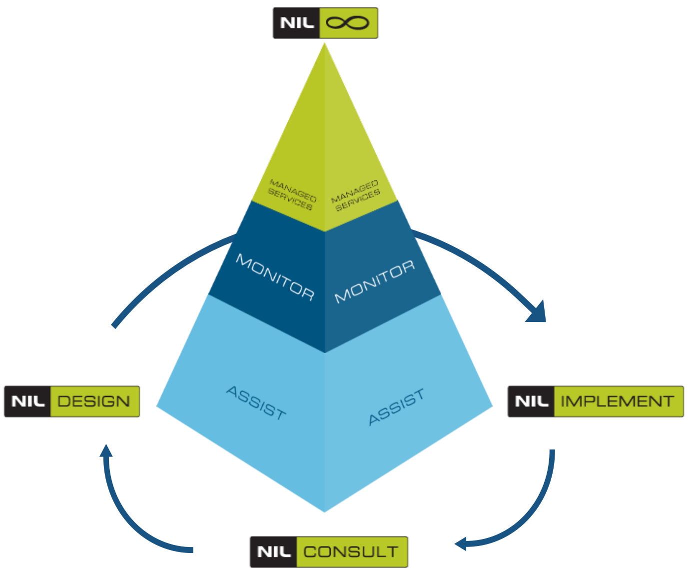
Need help?

ipSpace.net Consulting

- **ExpertExpress** for quick discussions, reviews or second opinions
- Short on-site technology, architecture or design workshops

NIL’s Professional/Learning Services

- In-depth design/deployment projects
- Data Center-, virtualization- and cloud-related training
- Details: www.nil.com, flipit.nil.com



Questions?

Send them to ip@ipSpace.net or @ioshints